(2)

ESTIMATION IN THE MIXTURE TRANSITION DISTRIBUTION MODEL FOR HIGH ORDER MARKOV CHAINS

by

Simon Tavaré
Adrian E. Raftery

TECHNICAL REPORT No. 211

June 1991

Department of Statistics, GN-22

University of Washington

Seattle, Washington 98195 USA

Estimation in the Mixture Transition Distribution Model for High Order Markov Chains

Simon Tavaré
University of Southern California *

Adrian E. Raftery
University of Washington

Submitted to *Applied Statistics*
6/7/91

Abstract

The mixture transition distribution (MTD) model was introduced by Raftery (1985) as a parsimonious model for high-order Markov chains. It is flexible, can represent a wide range of dependence patterns, can be physically motivated, fits data well, and appears to be a discrete-valued analogue for the class of autoregressive time series models. However, estimation has presented difficulties because the parameter space is highly nonconvex, being defined by a large number of nonlinear constraints.

Here we propose an efficient computational algorithm for maximum likelihood estimation which is based on a way of reducing the large number of constraints. This also allows more structured versions of the model, for example those involving structural zeros, to be fit quite easily. A way of fitting the model using GLIM is also discussed.

The algorithm is applied to a sequence of wind directions, and also to two sequences of DNA bases from introns from genes of the mouse. In each case, the MTD model fits better than the conventional Markov chain model.

1 Introduction

There are many examples in which we would like to fit high-order Markovian models to discrete data. However, in the conventional parametrisation of such processes the number of parameters increases geometrically with the order, so that parsimony is effectively lost. In this

*Simon Tavaré is Professor, Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113. Adrian E. Raftery is Professor, Departments of Statistics and Sociology, GN-22, University of Washington, Seattle, WA 98195. The authors are grateful to Brian Francis for helpful discussions about GLIM, to Jia Ye for help with the computations, to Peter Alfeld for several discussions about hashing, and to Ben Koop for sharing his data prior to publication. Simon Tavaré was supported in part by NSF grants DMS88-03284 and DMS90-05833 and NIH grant GM 41746. Adrian E. Raftery was supported by ONR Contract N-00014-88-K-0265.

paper, we describe some computational algorithms for fitting a parsimonious autoregressive-like Markov model known as the *Mixture Transition Distribution* (MTD) model, and we illustrate their use with some examples.

The MTD model was introduced by Raftery (1985a,b) and is defined as follows. Let $\{X_t : t = 0, 1, \ldots\}$ be a time homogeneous lth order Markov chain on a finite set of m states (here labelled 1,2,... ,m), and let the transition probabilities be

$$p(i_0 \,|\, i_1, \ldots, i_l) = P\,(X_{t+l} = i_0 \,|\, X_{t+l-1} = i_1, \ldots, X_t = i_l),\ t = 0, 1, \ldots. \tag{1}$$

There are $(m-1)\,m^l$ independent parameters in equation (1). Raftery's model provides, for $l > 1$, a useful parameter reduction in (1) by supposing that

$$p(i_0 \,|\, i_1, \ldots, i_l) = \sum_{j=1} \lambda_j\, q(i_0 \,|\, i_j), \tag{2}$$

where $Q = \{q(i \,|\, j)\}$ is a *column* stochastic matrix satisfying

$$q(i \,|\, j) \geq 0 \ \text{ and } \ \sum_{r=1}^{m} q(r \,|\, j) = 1,\ j = 1, \ldots, m, \tag{3}$$

and

$$\lambda_1 + \cdots + \lambda_l = 1. \tag{4}$$

Note that the number of independent parameters is now $m(m-1) + l - 1$, increasing only linearly in l. For example, when there are $m = 4$ states, the number of parameters for a second-order ($l = 2$) chain is 13 in the MTD model as against 48 in the usual second-order Markov chain model. To ensure that the transition probabilities are properly defined, we also require

$$\sum_{j=1}^{l} \lambda_j\, q(i_0 \,|\, i_j) \geq 0 \text{ for all } i_0, \ldots, i_l. \tag{5}$$

Notice that when (3), (4) and (5) are satisfied, $p(i_0 \,|\, i_1, \ldots, i_l) \leq 1$ for all $i_0, i_1, \ldots, i_l$.

The MTD model is so called because the conditional probabilities in (2) are linear combinations of contributions from the past. It is analogous to the AR(l) model in that one extra parameter is added to the model for each extra lag and that the lagged bivariate distributions satisfy a system of matrix equations similar to the Yule-Walker equations. In some situations it has a direct physical interpretation in terms of the probability of returning to past states, or states close to them.

We have found the MTD model useful in practice, but it is not easily fitted because of the non-linear nature of (2) and the constraints in (5). From an algorithmic point of view,

there are really two special cases of the MTD model, these being determined by what is assumed (in addition to (3) and (4)) about the $\{\lambda_i\}$. With the positivity assumption

$$\lambda_i \geq 0, i = 1, \ldots, l, \tag{6}$$

the inequality in (5) is automatically satisfied. The examples given in Raftery (1985a) satisfied (6), but several of the data sets we have analysed do not. One of these is described later in the paper. Without this positivity assumption, equation (5) comes into play in a crucial way, and computationally it becomes very important to be able to reduce the large number of constraints that are operating there. In section 2 we show how this can be done and in section 3 we give several examples.

2 Parameter Estimation

2.1 Reducing the Number of Constraints

We saw in section 1 that the general MTD model must satisfy the $m^l(m-1)$ constraints in (5). For example, in the four-state second-order case, i.e. $m = 4$ and $l = 2$, the number of constraints is 48, so that the resulting constrained numerical optimization problem is computationally demanding. The following result, which reduces the effective number of constraints in (5) to m, is at the heart of our fitting algorithm:

Proposition 1 *Let* $T = \sum_{i:\lambda_i \geq 0} \lambda_i$, *and define* $q_-(i) = \min_{1 \leq j \leq m} q(i \mid j)$ *and* $q_+(i) = \max_{1 \leq j \leq m} q(i \mid j)$. *Then* $\sum_{j=1}^{l} \lambda_j q(i \mid i_j) \geq 0$ *for all* $i, i_1, \ldots, i_l$ *if and only if*

$$T q_-(i) + (1 - T) q_+(i) \geq 0 \text{ for all } i. \tag{7}$$

Proof If (7) holds, then

$$\begin{aligned} \sum_{j=1}^{l} \lambda_j q(i \mid i_j) &= \sum_{j:\lambda_j \geq 0} \lambda_j q(i \mid i_j) + \sum_{j:\lambda_j < 0} \lambda_j q(i \mid i_j) \\ &\geq \sum_{j:\lambda_j \geq 0} \lambda_j q_-(i) + \sum_{j:\lambda_j < 0} \lambda_j q_+(i) \\ &= T q_-(i) + (1 - T) q_+(i) \\ &\geq 0. \end{aligned}$$

Conversely, assume that $q_-(i) = q(i \mid p_0)$ and $q_+(i) = q(i \mid p_1)$. Then

$$\begin{aligned} T q_-(i) + (1 - T) q_+(i) &= \sum_{j:\lambda_j \geq 0} \lambda_j q(i \mid p_0) \\ &\quad + \sum_{j:\lambda_j < 0} \lambda_j q(i \mid p_1) \\ &\geq 0. \end{aligned}$$

A-1

2.2 Maximum Likelihood Estimation

The parameters Q and $\{\lambda_i\}$ can be fitted by maximum likelihood by maximising the log-likelihood

$$\log L = \sum n(i_0, i_1, \ldots, i_l) \ \log p(i_0 \,|\, i_1, \ldots, i_l), \tag{8}$$

where $n(i_0, i_1, \ldots, i_l)$ is the number of times the sequence $i_l \rightarrow i_{l-1} \rightarrow \cdots \rightarrow i_0$ occurs in the data, $p(i_0 \,|\, i_1, \ldots, i_l)$ is given by (2), and the sum is over all $i_0, i_1, \ldots, i_l$ with $n(i_0, i_1, \ldots, i_l) > 0$. The maximisation is subject to the constraints (3), (4) and (7).

While there are several numerical approaches that might be taken to this problem, we found that direct maximisation of (8) was effective. We used the sequential quadratic programming algorithm implemented as E04UCF in Mark 13 of NAG (Numerical Algorithms Group, 1988). Although derivatives of the objective function and the constraint functions may be calculated, we found that approximating these by finite differences was effective.

One troublesome part of the algorithm involves the storage and recovery of the counts $n(i_0, i_1, \ldots, i_l)$. For models with high values of l or m, the number m^{l+1} of potential patterns can be extremely large. We proceeded by labelling a pattern $\mathbf{i} = i_0, i_1, \ldots, i_l$ by the number

$$i = 1 + \sum_{j=0}^{l} (i_j - 1)\, m^{l+1-j}.$$

If the number of possible patterns was sufficiently small, we stored the whole (now one-dimensional) array of counts. On the other hand, the maximum number N, say, of patterns that can be observed in the data is a little less than the length of the observed time series, so in cases where m^{l+1} is very large compared to N, we calculated and stored the counts using a simple hashing algorithm with a vector of length approximately N.

The types of data we have analysed have led to several useful additional features of the programs, namely:

1. The positive λ model satisfying (6) and the more general model satisfying (7) may be fitted separately.

2. Structural zeroes in the Q matrix may be handled directly. Example 3.2 in Raftery (1985a) is of this type.

3. Predetermined λ values may also be set to zero, corresponding to the omission of those terms.

4. Fixed or random starts for the parameters in the iterative scheme are allowed. In the first instance, Q is estimated by the usual first-order transition matrix, and the λ_i are equal. In the second instance, random Q and λ_i are used, subject to (3) and (4). This facility is particularly useful in the iterative algorithm for determining whether a local or potentially global maximum has been reached. While we have no formal proof that a unique maximum exists, numerical evidence with some 20 data sets suggests that it does. A more formal assessment of this might follow Finch et al. (1989).

2.3 Minimum χ^2 estimation

As an alternative to maximum likelihood, we have also used minimum χ^2 estimation. The aim is to find Q and $\{\lambda_i\}$ that minimise

$$X^2 = \sum \frac{(n(i_0,\ldots,i_l) - e(i_0\ldots,i_l))^2}{e(i_0,\ldots,i_l)},$$

where

$$e(i_0,\ldots,i_l) = n(+,i_1,\ldots,i_l)\, p(i_0 \mid i_1,\ldots,i_l),$$

and $+$ denotes summation over that index. The sum is over all $n(i_0,i_1,\ldots,i_l)$ for which $n(+,i_1,\ldots,i_l) > 0$, and the constraints in (3), (4) and (7) apply. This is a useful alternative, since the fitted counts e from the optimisation are natural candidates as measures of goodness of fit of the model. Kwok (1988) and Li and Kwok (1989) have shown that in some special cases of the MTD model, the minimum χ^2 estimator has lower bias than the maximum likelihood estimator but about the same variance, and hence lower overall mean squared error.

The asymptotic theory of X^2 when the parameters have been estimated by Maximum Likelihood is given in Billingsley (1961), where it is shown that X^2 has approximately a χ^2 distribution when the chain really is of order l. The asymptotic behavior of X^2 in the present context is the same.

We also experimented with different numerical algorithms for this problem, essentially based on knowledge of derivatives for the constraint functions. Once more the direct approach seems the easiest, using NAG algorithm E04UCF again.

Further details of the programs, and copies of the code, may be obtained from the authors upon request.

2.4 GLIM analysis of two-state models

When $m = 2$, the MTD model may be fitted using an iterative procedure in GLIM (Baker, 1986). Focus for the moment on the case $l = 2$, and write $\lambda_1 = \lambda, \lambda_2 = 1 - \lambda$. The log-likelihood (8) may be written

$$\log L = \sum_{i_1,i_2} \left(\sum_{i=1}^{2} n(i, i_1, i_2) \log p(i \,|\, i_1, i_2) \right). \tag{9}$$

For each i_1, i_2, the inner term in (9) is (essentially) a binomial log-likelihood for $n(+, i_1, i_2)$ trials and success probability $p(1 \,|\, i_1, i_2)$, where

$$p(1 \,|\, i_1, i_2) = \begin{cases} q(1 \,|\, 1) & i_1 = 1, i_2 = 1 \\ \lambda q(1 \,|\, 1) + (1 - \lambda) q(1 \,|\, 2) & i_1 = 1, i_2 = 2 \\ (1 - \lambda) q(1 \,|\, 1) + \lambda q(1 \,|\, 2) & i_1 = 2, i_2 = 1 \\ q(1 \,|\, 2) & i_1 = 2, i_2 = 2. \end{cases} \tag{10}$$

If λ is known, then (10) shows that the $p(1 \,|\, i_1, i_2)$ are linear in the parameters $q(1 \,|\, 1)$ and $q(1 \,|\, 2)$, $q(1 \,|\, 1)$ being the coefficient of the covariate $\mathbf{x}_1^T = (1, \lambda, 1 - \lambda, 0)$ and $q(1 \,|\, 2)$ the coefficient of the covariate $\mathbf{x}_2^T = (0, 1 - \lambda, \lambda, 1)$. Thus $q(1 \,|\, 1)$ and $q(1 \,|\, 2)$ (and so Q) may be estimated using binomial error, identity link, no intercept and covariates $\mathbf{x}_1$ and $\mathbf{x}_2$.

On the other hand, if $q(1 \,|\, 1)$ and $q(1 \,|\, 2)$ are assumed known, then (10) shows that $p(1 \,|\, i_1, i_2)$ is linear in λ; λ is the coefficient of the covariate $\mathbf{x}_3^T = (0, q(1 \,|\, 1) - q(1 \,|\, 2), q(1 \,|\, 2) - q(1 \,|\, 1), 0)$ and the offset is $\mathbf{x}_4^T = (q(1 \,|\, 1), q(1 \,|\, 2), q(1 \,|\, 1), q(1 \,|\, 2))$. Thus λ may be estimated using binomial error, identity link, no intercept, covariate $\mathbf{x}_3$, and offset $\mathbf{x}_4$. This leads to a simple recursive scheme for estimating the parameters, reminiscent of the iterative algorithms used in survival analysis; cf. Aitken *et al.* (1989), Chapter 6.

The generalisation to $l > 2$ is almost immediate from the form of (10). The number of covariates for the first stage remains 2, the elements of $\mathbf{x}_1$ being replaced by $\sum_{j:i_j=1} \lambda_j$. For the second stage the number of covariates is $l - 1$. It does not, however, seem to be simple to generalise this scheme to the case $m > 2$.

2.5 Model comparison

In order to compare the rival, non-nested, models in the examples that follow, we would ideally like to compute the posterior probability of each model under a range of plausible prior distributions for the parameters. The use of successive significance tests seems less satisfactory because many of the comparisons involve non-nested models and because the

use of multiple tests make the properties of the overall procedure hard to assess. We do not adopt information criteria for the selection of a single model, because conditioning on a single selected model ignores model uncertainty and so overstates our knowledge.

Here we use the approximate result that if we are comparing two models, M_0 and M_1, then the Bayes factor, or ratio of posterior to prior odds, B_{01}, for M_0 against M_1 satisfies

$$-2\log B_{01} = BIC_0 - BIC_1. \tag{11}$$

In (11), $BIC_i = -2\log L_i + k_i \log n$, where L_i is the maximized likelihood and k_i is the number of independent parameters in the model $M_i (i = 0, 1)$. Although this has not been formally proved for the MTD model, it has been established for independent exponential family observations by Schwarz (1978), for the usual Markov chain by Katz (1981) and for log-linear models of contingency tables by Raftery (1986a). The MTD model appears to satisfy regularity conditions that would permit the adaptation to it of proofs for other cases. If (11) is always some baseline model such as the independence model and (11) is calculated for each model of interest M_1, then the resulting Bayes factors readily yield the posterior probability of each of the models of interest (Raftery, 1988). The rules of thumb of Jeffreys (1961, Appendix B) suggest that such a comparison should not be regarded as decisive unless the difference in BIC values is at least about 10.

Model comparisons based on posterior probabilities can yield results different from those based on significance tests. This is especially so with large samples, including some that we analyse here. In such cases significance tests at fixed significance levels often reject null hypotheses more easily; an example with $n \doteq 110{,}000$ was discussed by Raftery (1986b). This is related to the "conflict between significance and P values" discussed by Berger and Sellke (1987). Alternatively, basing model comparisons on Bayes factors may be viewed as an automatic, decision-theoretic, way of setting significance levels so as to balance power and significance.

Our code produces Pearson residuals which can be used to suggest ways in which the model could be improved. New models suggested by such a process can be compared with the other models under consideration also using approximate Bayes factors.

3 Wind direction data

Raftery *et al.* (1982) and Haslett and Raftery (1989) describe a data set that includes hourly observations of wind directions at a meteorological station at Roche's Point, Ireland. The

data that we analyse here began at 1 am on January 1, 1961 and ran for almost 9 years. There are 77,155 observations in all. The original data were recorded as 0 for no wind, and then in 10 degree units from due North, for a total of 37 states.

One aim here is to predict wind speeds and directions so as to control the wind turbine generators making up a wind farm, and to manage the electric power supply. Wind turbines should be oriented in such a way as to derive the most energy from the wind, so that their current best orientation is a function of future as well as current wind direction. Predicting output from a variable energy source such as wind is important so that the need for power from other, more stable, sources such as oil can be anticipated.

Figure 1 provides a histogram of the distribution of wind directions for all nine years combined, together with separate histograms for each of the 8 complete years in the data set. Broadly speaking, these annual histograms are rather similar and show natural modes in the data that are preserved from year to year. Based on these results, we chose to recode the data into five categories: 0, 6-14, 15-23, 24-32, and 33-5; these are labelled 1 to 5 respectively in what follows.

Insert Figure 1 about here

As might have been anticipated, there are some inhomogeneities in the distribution of wind directions which are revealed when the data are analysed in separate months. See Figure 2.

Insert Figure 2 about here

These distributions are rather similar for the months of November through April. We therefore chose the months November through April as a period in which wind directions might be modelled by a stationary MTD model. The data analyzed below come from the period November 1961 to April 1962, providing a total of 4344 consecutive hourly observations.

Insert Table 1 about here

Table 1 gives the BIC values for the full Markov model, and the MTD model. The order is estimated to be 7. The estimated parameters are $\hat{\lambda}_1 = 0.591, \hat{\lambda}_2 = .237, \hat{\lambda}_3 = .076, \hat{\lambda}_4 = .018, \hat{\lambda}_5 = .024, \hat{\lambda}_6 = .024, \hat{\lambda}_7 = .031$, while

$$\hat{Q} = \begin{pmatrix} .65 & .01 & .01 & .01 & .01 \\ .09 & .95 & .01 & .00 & .03 \\ .14 & .02 & .92 & .04 & .00 \\ .05 & .00 & .06 & .92 & .04 \\ .08 & .03 & .00 & .03 & .91 \end{pmatrix}.$$

The estimated $\hat{\lambda}_j$'s are positive, indicate that the most recent observations are the most important, and that the current observation tends to be close to the immediately preceding ones, as we would expect. The $\hat{Q}$ matrix indicates the process to be smooth, with the probability of staying in the same state being 0.91 or greater whenever there is any wind, and the probability of the direction changing by more than one state being very small.

We note that $\hat{\lambda}_7$ is larger than $\hat{\lambda}_4$, $\hat{\lambda}_5$ and $\hat{\lambda}_6$, probably due to the fact that the λ_7 term is capturing the small residual dependence on $X_{t-8}, X_{t-9}, \ldots$, as well as the dependence on X_{t-7} itself. This suggested that we fit another MTD(7) model, with the constraint that $\lambda_4 = \lambda_5 = \lambda_6 = 0$. As we discussed in section 2.2, this is easy to do using our algorithm.

The resulting BIC value was 4530.4, making this quite clearly the best model considered. The $\hat{Q}$ matrix was almost unchanged, while $\hat{\lambda}_1 = .598$, $\hat{\lambda}_2 = .245$, $\hat{\lambda}_3 = .100$, $\hat{\lambda}_7 = .057$. This is not very different from the full MTD(7) model, but it seems to summarize the dependence in a more parsimonious way.

4 The analysis of intron sequences

The area of biomolecular sequence comparison has provided statisticians with a wealth of novel problems. As an example, descriptive statistics on DNA composition have proved useful in the search for coding regions and introns, the statistical assessment of sequence similarity, and the analysis of repeated motifs that may be of biological significance. Similarly, statistical analysis of protein sequences of known three-dimensional structure has been used to infer potential folding patterns of other proteins. It is not our purpose here to describe these areas in any detail; rather, we refer the reader to the recent review article by Curnow and Kirkwood (1989) and the books edited by Waterman (1988) and Doolittle (1990) for good introductions to this important field. In this section, we will focus on just one example from the area of DNA sequence analysis to which the MTD model might be applied.

The statistical significance of repeated patterns in a DNA sequence must, of course, be measured against the background stochastic structure of the sequence itself. Among possible models for this structure are Markov chains, which might describe the DNA sequence in

terms of its nucleotide composition (that is, as a string of letters from a four-letter alphabet, $\{A, C, G, T\}$). There are several other alphabets of biological interest such as the purine-pyrimidine alphabet in which each base in the sequence is coded as either purine ($\{A, G\}$) or pyrimidine ($\{C, T\}$). For example, Blaisdell (1983) reported that, relative to a model of independent bases, non-coding sequences (such as introns) generally contain a shortage of runs of length 1 and 2 of purines and pyrimidines, and an excess of long runs of them; see also Karlin, Ost and Blaisdell (1988). In this section, we describe an exploratory analysis of two different DNA sequences from introns in certain mouse genes.

4.1 The mouse T-cell receptor α/δ locus

The first example is an analysis of part of the mouse T-cell receptor α/δ locus (Wilson *et al.*, 1991; Koop *et al.*, 11991). This region is 94,647 bases in length. It comprises over 50 introns and 50 exons; the exons comprise just 6% of the sequence. The particular sequence we have analyzed is the intron prior to joining gene segment J50 (Koop *et al.*, 1991). It starts 5′ to exon 1 of Vδ5, and ends three bases before the recombination signal 5′ to J50. The sequence is 5,778 bases in length.

A preliminary analysis shows that the sequence is clearly first-order when analyzed in the four-letter alphabet $\{A, G, C, T\}$; see Table 2. The MTD model provides no improvement on this fit. The estimated transition matrix $\hat{Q}$ is

		last base			
		A	G	C	T
	A	.31	.29	.32	.18
next	G	.27	.27	.04	.29
base	C	.20	.22	.27	.26
	T	.22	.21	.37	.27

Insert Table 2 about here

Recall that a Markov chain with transition matrix Q is (strongly) lumpable with respect to a partition $P_1, \ldots, P_r$ of the state space if, and only if, for $1 \leq i, j \leq r$, $\sum_{l \in P_j} q_{lk}$ has the same value for each $k \in P_i$ (Kemeny and Snell, 1960). The lumpability condition ensures that the lumped process is also Markovian, no matter what the initial distribution of the original process might be. An examination of the matrix $\hat{Q}$ above indicates that we might simplify the stochastic description of this intron by lumping the states A and G into a single state A/G that denotes purine. A formal test of this lumpability hypothesis may be found

in Thomas and Barr (1977). The new alphabet is $\{A/G, C, T\}$. The BIC analysis of this new sequence is presented in Table 2. As would be expected, among the fully-parametrized Markov models a first order chain provides the most parsimonious description. Its estimated transition matrix is

		last base		
		A/G	C	T
next	A/G	.57	.36	.47
base	G	.21	.27	.26
	C	.22	.37	.27

However, it can be seen from Table 2 that a second order MTD model provides a better description. In this case, $\hat{\lambda}_1 = 0.71, \hat{\lambda}_2 = 0.29$, and the estimated transition matrix $\hat{Q}$ is

		last base		
		A/G	C	T
next	A/G	.60	.33	.44
base	G	.20	.28	.26
	C	.20	.39	.29

Finally, we analyze the purine-pyrimidine alphabet $\{A/G, C/T\}$. From the previous discussion, it seems clear that the original sequence is not lumpable with respect to the purine-pyrimidine partition of the states. The purine-pyrimidine sequence may not be Markovian (or even homogeneous, unless we assume that the original chain was stationary). Fitting Markovian models to such data provides an exploratory approach to approximating the stochastic structure of a complicated process by simpler ones. One might expect this more complicated structure to be reflected in a higher estimated order of dependence in the Markovian approximation. This is indeed the case here, as the results in Table 2 verify. The purine-pyrimidine chain is approximated by a second-order Markov chain. The MTD model offers no further improvement in this case.

The discussion of lumpability provides an indication of the greatest extent to which the states of the chain can be aggregated without losing important structure. Here, this seems to be the three-state case analyzed above. Thus, in a sense, the second-order MTD model for the three-state case provides the most parsimonious available representation of the data within the class of Markov chain models discussed.

4.2 The mouse αA-crystallin gene

Avery (1987) examined the Markovian structure of introns from several other genes in mouse, in order to determine whether certain short DNA sequences occurred more often than would

be expected by chance. Here we will analyse the introns from the mouse αA-crystallin gene, further details of which may be found in Avery's paper. The sequence analyzed here comprises two introns, of total length 1307 bases.

Following the style of analysis of the previous example, we see first that the sequence of bases with alphabet $\{A, C, G, T\}$ is clearly indicated to be of order 1; see Table 3. The estimated transition matrix $\hat{Q}$ is given by

		last base			
		A	G	C	T
	A	.23	.23	.30	.19
next	G	.34	.32	.06	.30
base	C	.25	.27	.34	.28
	T	.18	.19	.30	.23

Insert Table 3 about here

Notice that this transition matrix is qualitatively rather similar to the corresponding matrix for the T-cell receptor intron discussed in the previous section. In particular, this sequence is also (approximately) lumpable with respect to the partition $\{A/G, C, T\}$. The results are again consistent with the previous example, in that among the fully-parametrized Markov models, the first order model provides the best description. However, a more parsimonious description is provided by an MTD(2) model in which $\hat{\lambda}_1 = 2.46, \hat{\lambda}_2 = -1.46$, and the estimated transition matrix $\hat{Q}$ is

		last base		
		A/G	C	T
next	A/G	.52	.43	.49
base	G	.27	.32	.29
	C	.21	.25	.22

Note that in this example the likelihood is maximized by some negative values of the λ_i. The constraints are maintained by using the method outlined in Proposition 1.

Finally, the analysis of the collapsed chain in its purine-pyrimidine alphabet is given in Table 3. The odds for the data being generated by a second-order MTD model as against a first-order Markov chain are about 2:1, by (11). This does provide some evidence for the chain being of order two, although in the words of Jeffreys (1961) it is "not worth more than a bare mention". (The standard likelihood ratio test statistic is 8.7 with one degree of

freedom, and the corresponding P-value from the asymptotic χ^2 distribution is about 0.003. Thus the approximate Bayes factor and the approximate P-value point in the same direction but, as usual, the P-value suggests stronger evidence for the larger model.) The parameter estimates from the GLIM algorithm described in section 2.4 are identical, and the minimum χ^2 estimates are essentially the same.

In this example, the structure of the purine-pyrimidine sequence is captured by the MTD model, rather than by the fully parametrized Markov model that was required for the earlier intron sequence. The parameters are $\hat{\lambda}_1 = 2.19, \hat{\lambda}_2 = -1.19$ and the estimated transition matrix $\hat{Q}$ is

		last base	
		A/G	C/T
next	A/G	.52	.45
base	G	.48	.55

4.3 Comments

In these examples the emphasis is on model fitting to find (or approximate) structure, rather than for prediction. We have seen that the MTD(2) model provides a good description of the two intron sequences when they are coded in the $\{A/G, C, T\}$ alphabet. Some other sequence analysis examples in which the MTD model has been applied appear in Tavaré and Giddings (1988). While Markov models are a useful first step in this context, their validity is often questionable because of possible inhomogeneities in the sequence. This inhomogeneity is particularly pronounced in coding regions (exons), where it is well known that the three codon positions exhibit markedly different behavior. To analyse such regions, more sophisticated non-homogeneous Markov models may be required. Some of these are described, for example, by Tavaré and Song (1989) and Watterson (1991).

5 Discussion

Various generalisations of the MTD model have been proposed. Raftery (1985a,b) proposed ways of modelling the case where $m = \infty$, such as when the observations are counts. Adke and Deshmukh (1988) showed that asymptotic properties valid when m is finite also apply when $m = \infty$. It seems that our estimation method will work in that case also, provided that the (now doubly infinite) matrix Q is modelled parametrically. If $m = \infty$ and

$$\liminf_{i \to \infty} q(i|j) = 0 \quad \forall j, \tag{12}$$

then the constraints (5) are equivalent to the positivity assumption (6), and the computational problem is greatly simplified.

Mehran (1989) considered the infinite lag MTD model, $l = \infty$, where λ_j is a parametric function of j. Our method seems applicable in this case also, although calculating the likelihood, or the fitted values for minimum χ^2 estimation, seems difficult. It may be possible to model discrete-valued time series with the long-memory property using this approach, by setting the λ_j equal to the π-weights for the fractionally-differenced ARIMA (p, d, q) process. Various continuous-valued environmental time series such as wind speeds are of this kind (Haslett and Raftery, 1989), and it seems reasonable to suppose that some discrete-valued time series might have this property also.

Martin and Raftery (1987) and Adke and Deshmukh (1988) pointed out that the MTD model remains well-defined for arbitrary state spaces, which need not be finite, countable or even discrete. Equations (1) and (2) remain valid if p and q are interpreted as conditional densities, where q will usually have some parametric form. Le, Martin and Raftery (1990) have shown that this provides a framework for modelling bursts, outliers and flat stretches in continuous-valued time series, and also models time series that are well fitted by conventional Gaussian ARMA models. If condition (12) holds, then so does the positivity assumption (6). However, this is not always the case, even when the state space is continuous. For example, in continuous-valued directional time series, (12) does not necessarily hold. Craig (1989) has investigated MTD and other models for this situation, and has studied the consequences of (12) not holding. Breckling (1989) has also studied such time series.

6 References

ADKE, S.R. and DESHMUKH S.R. (1988) Limit distribution of a high order Markov chain. *J. R. Statist. Soc. B*, **50**, 105-108.

AITKIN, M., ANDERSON, D., FRANCIS, B. and HINDE, J. (1989) *Statistical modelling in GLIM*, Oxford Statistical Science Series, Volume 4. Clarendon Press, Oxford.

AVERY, P.J. (1987) The analysis of intron data and their use in the detection of short signals. *J. Mol. Evol.*, **26**, 335-340.

BAKER R.J. (1986) *The GLIM Release 3.77 Reference Guide*, Numerical Algorithms Group, Oxford.

BERGER, J.O. and SELLKE, T. (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.*, **82**, 112-122.

BILLINGSLEY, P. (1961) *Statistical Inferences for Markov Processes*, University of Chicago Press.

BLAISDELL, B.E. (1983) A prevalent persistent global nonrandomness that distinguishes coding and non-coding eucaryotic nuclear DNA sequences. *J. Mol. Evol.*, **19**, 122-133.

BRECKLING J. (1989) *The Analysis of Directional Time Series: Applications to Wind Speed.* Springer Lecture Notes in Statistics, 61. Springer, New York.

CRAIG, P. (1989) *Time Series Analysis of Directional Data.* Unpublished Ph.D. thesis, Department of Statistics, Trinity College, Dublin.

CURNOW, R.N. and KIRKWOOD, T.B.L. (1989) Statistical analysis of deoxyribonucleic acid sequence data – a review. *J. R. Statist. Soc. A*, **152**, 199-220.

DOOLITTLE, R.F. (1990) Editor, *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences.* Methods in Enzymology, Volume 183. Academic Press, San Diego.

FINCH, S.J., MENDELL, N.R. and THODE, H.C. Jr. (1989) Probabilistic measures of the adequacy of the search for a global maximum. *J.A.S.A.*, **84**, 1020-1023.

HASLETT J. and RAFTERY A.E. (1989) Space-time modelling with long-memory dependence: assessing Ireland's wind power resource. *Applied Statistics*, **38**, 1-50.

JEFFREYS, H. (1961) *Theory of Probability*, Third Edition, Oxford: Clarendon.

KARLIN, S., OST F. and BLAISDELL, B.E. (1988) Patterns in DNA and amino acid sequences and their statistical significance. *Mathematical Methods for DNA Sequences*, 133-157, edited by M.S. Waterman, CRC Press.

KATZ, R.W. (1981) On some criteria for estimating the order of a Markov chain. *Technometrics*, **23**, 243-249.

KEMENY, J.G. and SNELL, J.L. (1960) *Finite Markov Chains.* Van Nostrand.

KOOP, B.F., WILSON, R.K., WANG, K., VERNOOIJ, B., ZALLER, D., LAM, C., SETO, D. and HOOD, L. (1991) Organization, structure and function of the murine T-cell receptor Cα to Cδ region. In preparation.

KWOK, M.C.O. (1988) Some results on higher order Markov chain models. M. Phil. thesis, University of Hong Kong.

LE, N.D., MARTIN, R.D. and RAFTERY, A.E. (1990) Modeling outliers, bursts and flat stretches in time series using Mixture Transition Distribution (MTD) models. Technical Report no. 194, Department of Statistics, University of Washington.

LI, W.K. and KWOK, M.C.O. (1989) Some results on the estimation of a higher order Markov chain. Unpublished manuscript.

MARTIN R.D. and RAFTERY, A.E. (1987) Outliers, computation and non-Euclidean models. *J. Amer. Statist. Assoc.*, **82**, 1044-1050.

MEHRAN, F. (1989) Analysis of discrete longditudinal data: infinte lag Markov models. In *Statistical Data Analysis and Inference*, (Y. Dodge, ed.) 533-541. Elsevier.

NAG (1988) Fortran Library, Mark 13. Numerical Algorithms Group Ltd., Oxford, England.

RAFTERY, A.E. (1985a) A model for high-order Markov chains. *J. R. Statist. Soc. B*, **47**, 528-539.

RAFTERY, A.E. (1985b) A new model for discrete-valued time series: autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni*, **3-4**, 149-162.

RAFTERY, A.E. (1986a) A note on Bayes factors for log-linear contingency table models with vague prior information. *J. Royal Statist. Soc., B*, **48**, 249-250.

RAFTERY, A.E. (1986b) Choosing models for cross-classifications. *American Sociological Review*, **51**, 145-146.

RAFTERY, A.E. (1988) Approximate Bayes factors for generalized linear models. Technical Report no. 121, Department of Statistics, University of Washington.

RAFTERY A.E., HASLETT J. and McCOLL, E. (1982) Wind power: a space-time process? *Time series analysis: theory and practice* O.D. Anderson (ed.), 191-202. North-Holland Publishing Company.

SCHWARZ, G.(1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

TAVARÉ, S. and GIDDINGS, B.W. (1988) Some statistical aspects of the primary structure of nucleotide sequences. *Mathematical Methods for DNA Sequences*, 117-132, edited by M.S. Waterman, CRC Press, Boca Raton, FL.

TAVARÉ, S. and SONG, B. (1989) Codon preference and primary sequence structure in protein coding regions. *Bull. Math. Biol.*, **51**, 95-115.

THOMAS, M.U. and BARR, D.R. (1977) An approximate test for Markov chain lumpability. *J. Amer. Statist. Assoc.*, **72**, 175-179.

WATERMAN, M.S. (1988) Editor, *Mathematical Methods for DNA Sequences.* CRC Press, Boca Raton, FL.

WATTERSON, G.A. (1991) A stochastic analysis of three viral sequences. *Mol. Biol. Evol.*, submitted.

WILSON, R.K., KOOP, B.F., CHEN, C., HALLORAN, N., SCLAMMIS, R. and HOOD, L. (1991) Nucleotide sequence analysis of the 3' terminal region of the murine T-cell receptor α/δ chain locus. In preparation.

Caption for Figure 1

Histograms of wind directions in Roche's Point, Ireland, 1961-1969.
Analysis by year.
Wind directions are measured in units of 10° from North. 0 indicates no wind.

Caption for Figure 2

Histograms of wind directions in Roche's Point, Ireland, 1961-1969.
Analysis by month.
Wind directions are measured in units of 10° from North. 0 indicates no wind.

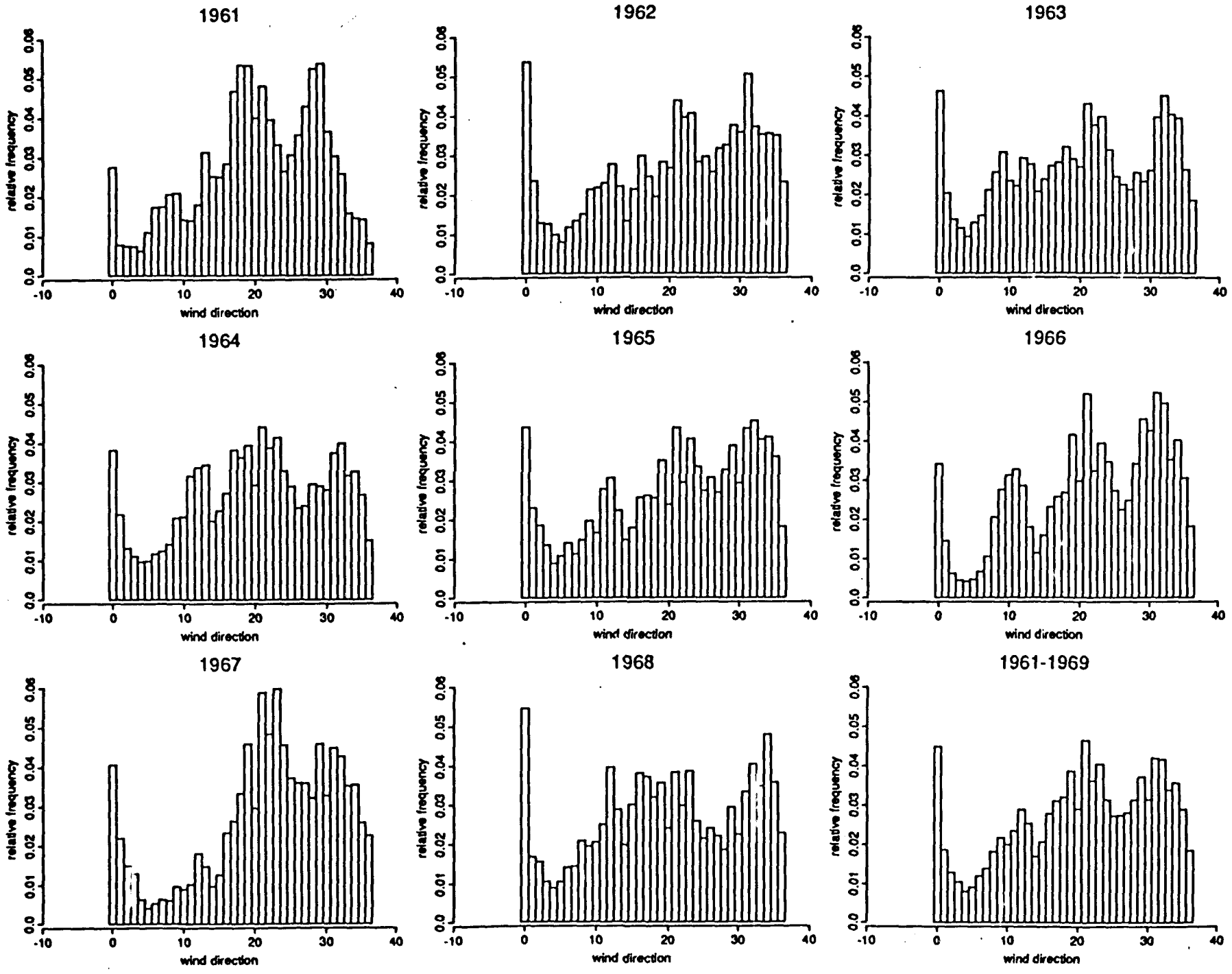
1961
1962
1963
1964
1965
1966
1967
1968
1961-1969
relative frequency
wind direction
0.0
0.01
0.02
0.03
0.04
0.05
0.06
-10
0
10
20
30
40

FIGURE 1

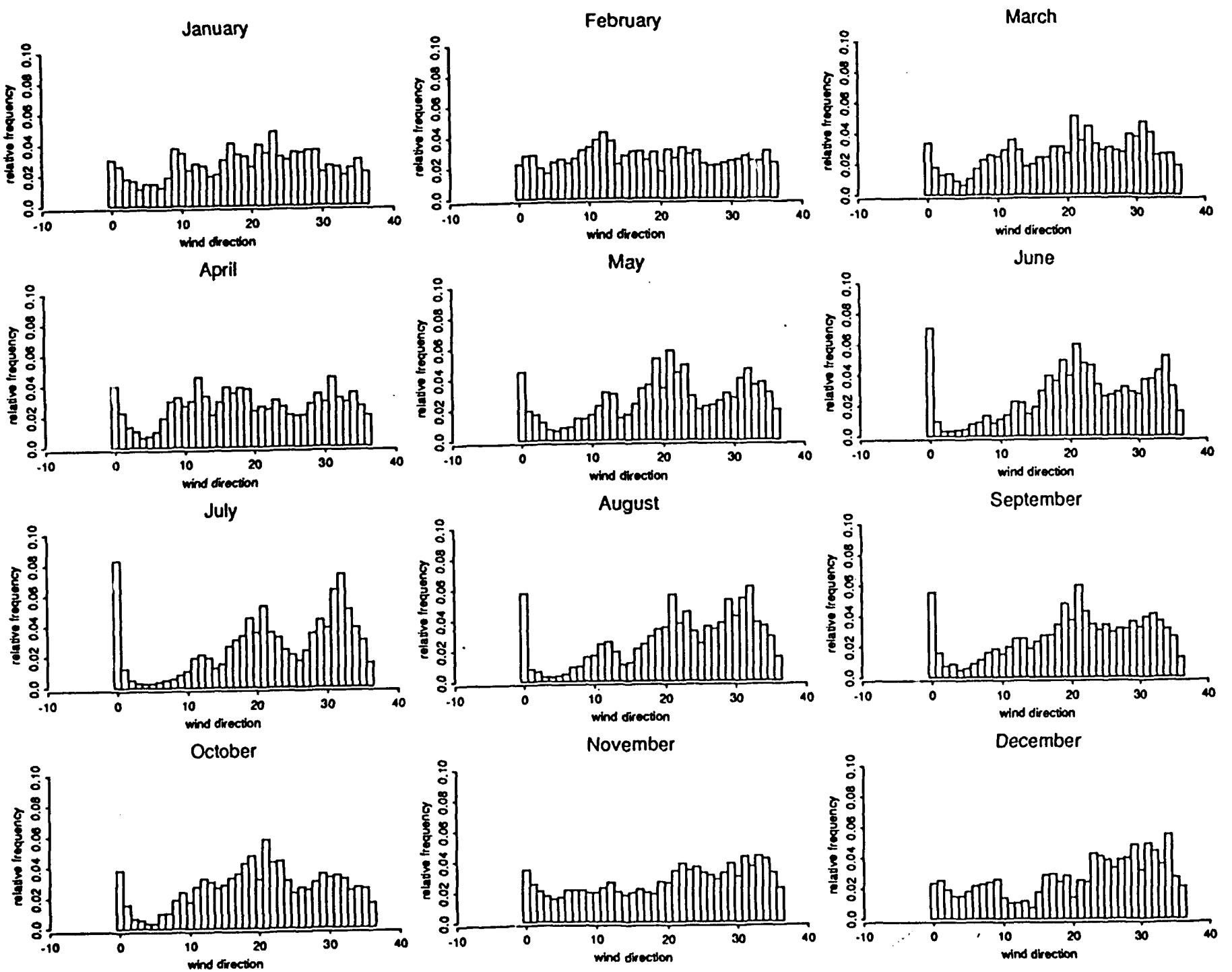
January
February
March
April
May
June
July
August
September
October
November
December
relative frequency
0.0 0.02 0.04 0.06 0.08 0.10
-10 0 10 20 30 40
wind direction

FIGURE 2

Table 1: Wind direction data from Roche's Point, Ireland[1].

Order, l	Number of parameters, k	BIC (full model)	Number of parameters, k	BIC (MTD model)
0	4	12,716.5		
1	20	5,085.7	20	5,085.7
2	100	5,198.8	21	4,646.4
3	500	8,243.3	22	4,569.5
4	2,500	24,674.7	23	4,557.5
5	-	-	24	4,544.7
6	-	-	25	4,539.8
7	-	-	26	4,538.8
8	-	-	27	4,540.8
9	-	-	28	4,540.4
10	-	-	29	4,545.4

[1] $n = 4,325$ observations starting at position 20 in the sequence.

Table 2: Intron from mouse T-cell receptor δ/α locus[1].

Order, l	Number of parameters, k	BIC (full model)	Number of parameters, k	BIC (MTD model)
	Alphabet:	A, C, G, T		
0	3	15,980.1		
1	12	15,549.3		
2	48	15,756.9	13	15,552.4
3	192	16,828.7	14	15,561.1
	Alphabet:	$A/G, C, T$		
0	2	12,042.0		
1	6	11,900.1		
2	18	11,939.7	7	11,885.5
3	54	12,204.7	8	11,890.2
	Alphabet:	$A/G, C/T$		
0	2	8,005.6		
1	4	7,881.8		
2	8	7,850.5	3	7,851.4
3	16	7,878.2	4	7,856.1

[1] $n = 5,369$ bases, starting at position 10 in the sequence.

Table 3: Introns from Mouse αA-crystallin gene[1]

Order, l	Number of parameters, k	BIC (full model)	Number of parameters, k	BIC (MTD model)
	Alphabet:	A, C, G, T		
0	3	3,620.8		
1	12	3,559.7		
2	48	3,758.8	13	3,566.1
3	192	4,542.8	14	3,572.8
	Alphabet:	$A/G, C, T$		
0	2	2,739.0		
1	6	2,728.7		
2	18	2,786.6	7	2,722.7
3	54	2,973.2	8	2,729.4
	Alphabet:	$A/G, C/T$		
0	1	1,810.9		
1	2	1,792.8		
2	4	1,798.1	3	1,791.3
3	8	1,813.8	4	1,797.1

[1] Two introns, n=1302 bases, starting at position 6 in the sequence.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 211	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Estimation in the Mixture Transition Distribution Model for High Order Markov Chains		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Simon Tavaré Adrian E. Raftery		8. CONTRACT OR GRANT NUMBER(s) N00014-88-K-0265
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of Washington Seattle, WA 98195		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-661-003
11. CONTROLLING OFFICE NAME AND ADDRESS —		12. REPORT DATE June 1991 13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) ONR Code N63374 1107 NE 45th Street Seattle, WA 98195		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See reverse side.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The mixture transition distribution (MTD) model was introduced by Raftery (1985) as a parsimonious model for high-order Markov chains. It is flexible, can represent a wide range of dependence patterns, can be physically motivated, fits data well, and appears to be a discrete-valued analogue for the class of autoregressive time series models. However, estimation has presented difficulties because the parameter space is highly nonconvex, being defined by a large number of nonlinear constraints.

Here we propose an efficient computational algorithm for maximum likelihood estimation which is based on a way of reducing the large number of constraints. This also allows more structured versions of the model, for example those involving structural zeros, to be fit quite easily. A way of fitting the model using GLIM is also discussed.

The algorithm is applied to a sequence of wind directions, and also to two sequences of DNA bases from introns from genes of the mouse. In each case, the MTD model fits better than the conventional Markov chain model.